Compositionally-warped additive modeling: COVID19の地理的要因分析への応用

2020年10月27日　統計数理研究所　オープンハウス

村上　大輔　　データ科学研究系　助教

【概要】

手法開発

地理空間データの大規模化・多様化が進む昨今、幅広いデータを柔軟かつ計算効率よくモデリングする方法が求められている。

この点を踏まえ、本研究では以下を満たす回帰を新規に開発する：

(a) 幅広い非ガウスデータがモデリングできる

(b) 幅広い効果が推定できる

(c) 計算効率が良い

Compositionally-warped Gaussian process (CWGP; Rois and Tober, 2019)を用いて(a)を、additive modelを用いて(b)を、データ行列のpre-conditioning (Murakami and Griffith, 2019)を用いて(c)を、それぞれ満たす。以上で開発する**Compositionally-warped additive mixed model (CAMM)**はRパッケージspmoranに近日中に実装予定

実データへの応用

COVID19陽性者数データ（都道府県別・日別）の要因分析に応用する。

【提案手法】

非ガウス分布に従う被説明変数y_iのための加法モデル（以下）を考える：

$$\varphi_{\boldsymbol{\Theta}}(y_i) = \sum_{k=1}^{K} x_{i,k}\beta_k + \sum_{l=1}^{L} f_l(z_l) + \varepsilon_k \qquad \varepsilon_k \sim N(0, \sigma^2)$$

変換関数$\varphi_{\boldsymbol{\Theta}}(y_i)$ → (a)への対処

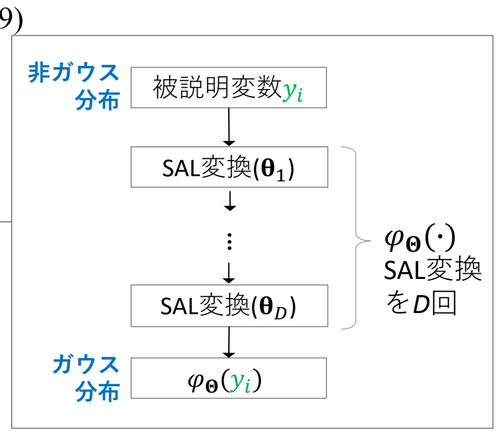
以下のSAL(Sinh-Arcsinh and Affine)変換をD回繰り返すことで、幅広い非ガウス分布がガウス分布に変換可能(Tober and Rois, 2019)

$$\varphi_d(y_i) = a_d + b_d \sinh(c_d \operatorname{arcsinh}(y_i) d_d)$$

そこで変換関数は下式(右図)で与える

$$\varphi_{\boldsymbol{\Theta}}(\cdot) = \varphi_D(\cdots \varphi_2(\varphi_1(\cdot))\cdots)$$

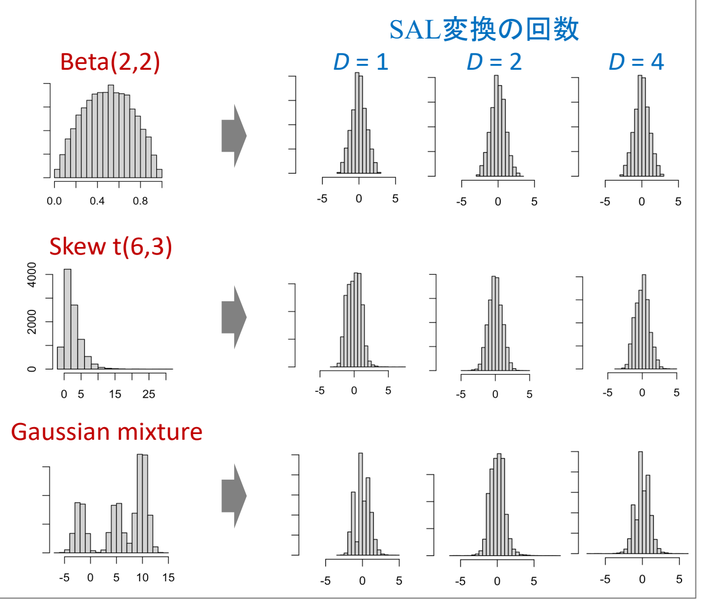
$\boldsymbol{\theta}_d \in \{a_d, b_d, c_d, d_d\}$: パラメータ
$\boldsymbol{\Theta} \in \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_D\}$



予備的検討

Beta分布, Skew t分布, 混合ガウス分布から生成した各変数y_iに上記変換$\varphi_{\boldsymbol{\Theta}}(\cdot)$を適用して、それらが適切にガウス分布に変換されるかを検証

適用結果（右図)より、変換回数Dが2以上であれば各分布がおおむねガウス分布に近似されていることが確認できる



変量効果を捉える関数$f_l(z_l)$ → (b)への対処

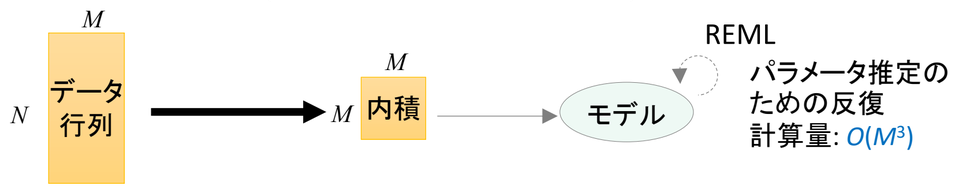
右のような幅広い効果を捉える加法モデルが研究されてきたことができる

(Umlauf et al., 2012)

- Nonlinear effects of continuous covariates: $f_j(\mathbf{z}) = f(z_1)$.
- Two-dimensional surfaces: $f_j(\mathbf{z}) = f(z_1, z_2)$.
- Spatially correlated effects: $f_j(\mathbf{z}) = f_{spat}(z_s)$.
- Varying coefficients: $f_j(\mathbf{z}) = z_1 f(z_2)$.
- Spatially varying effects: $f_j(\mathbf{z}) = z_1 f_{spat}(z_s)$
- Random intercepts with cluster index c: $f_j(\mathbf{z}) = \beta_c$.
- Random slopes with cluster index c: $f_j(\mathbf{z}) = z_1 \beta_c$.

高速なRestricted maximum lilelihood (REML)推定 → (c)への対処

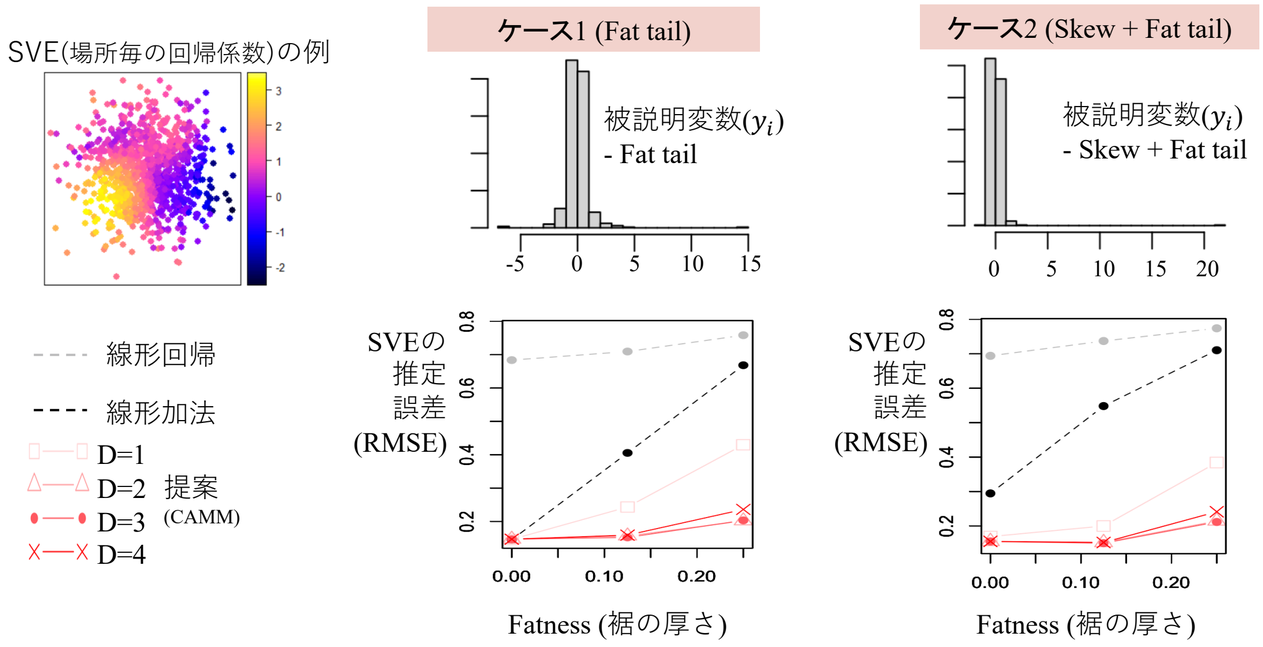
データ行列自体ではなく、同行列の内積を用いてパラメータ推定のための反復計算を行う → 高速推定可能(Murakami and Griffith, 2019)



参考文献： Rios and Tobar (2019). *Neural Netw*; Umlauf et al. (2012) J Stat Softw; Murakami and Griffith (2019) Spat Stat

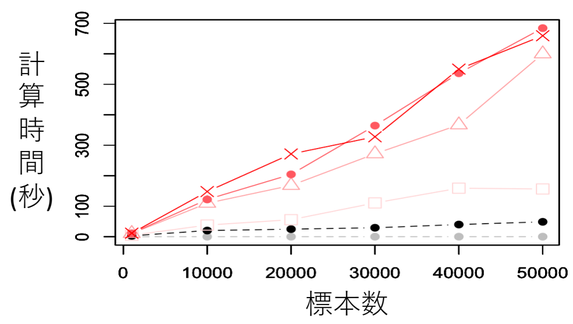
【Monte Carloシミュレーション】

- $f_l(z_l)$で捉える効果としてSpatially varying effects (SVE; 場所ごとの回帰係数)を想定。同回帰係数の推定精度をシミュレーション実験で評価
  - ✓被説明変数y_iの分布がFat tailなケース1と、SkewでFat tailなケース2の2条件で評価



結果

- 線形加法モデルの推定精度は非ガウスデータの場合に大幅に悪化
- 提案モデルの精度は分布によらず良好
  (特に変換回数Dが2または3の場合)
- 計算効率の良さも確認（右図）



【COVID19の地理的要因分析】

概要

- 提案手法をジャグジャパン社公開の陽性者数（都道府県・日別）データの解析に応用する。なお滞留人流はモバイル空間統計/ドコモ・インサイトマーケティング提供で推計しています。モバイル空間統計はＮＴＴドコモの登録商標です。

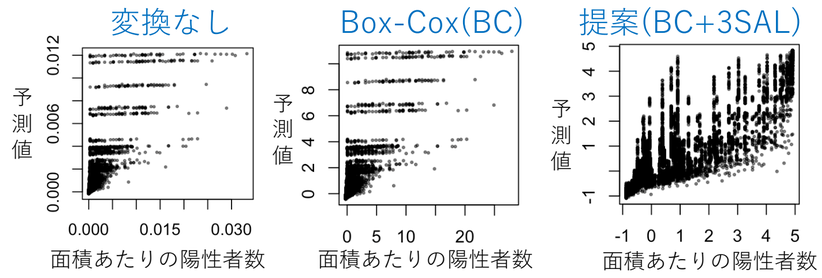
- 被説明変数(y_i)
  - ✓面積あたりの陽性者数 (年代別・日別・都道府県別; 3/1～6/3; N = 45,632)
- 説明変数($x_{i,k}$ またはz_l)
  - ✓滞留人口(住宅地)
  - ✓滞留人口(商業地)/滞留人口(住宅地)
  - ✓平均所得
  - ✓週
  - ✓曜日
  - ✓年齢
  - ✓都道府県
  - ✓年齢×都道府県

→ BICを改善する場合のみ、
- 回帰係数を都道府県毎に推定
- 回帰係数を説明変数の値毎に推定

→ グループ（変量）効果
- 強制投入

結果：モデルの精度

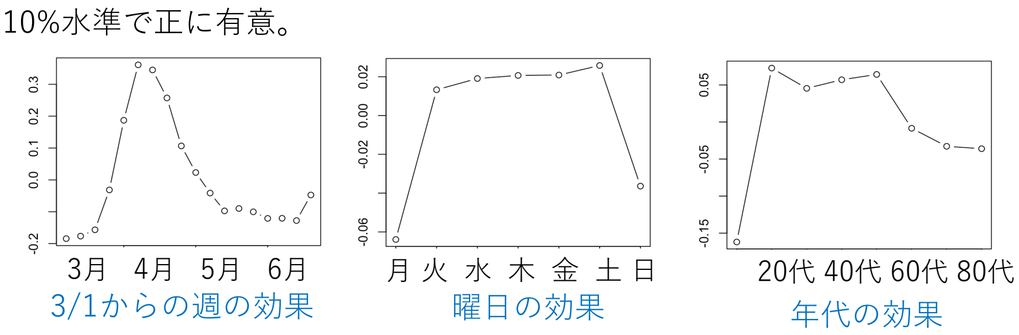
- 変換なしやBox-Cox変換に比べ大幅に精度が改善
- 計算効率の良さも確認



	変換なし	Box-Cox(BC)	提案(BC+3SAL)
BIC	-5.38 x 10^5	-5.46 x 10^5	-7.24 x 10^5
計算時間	353秒	485秒	483秒

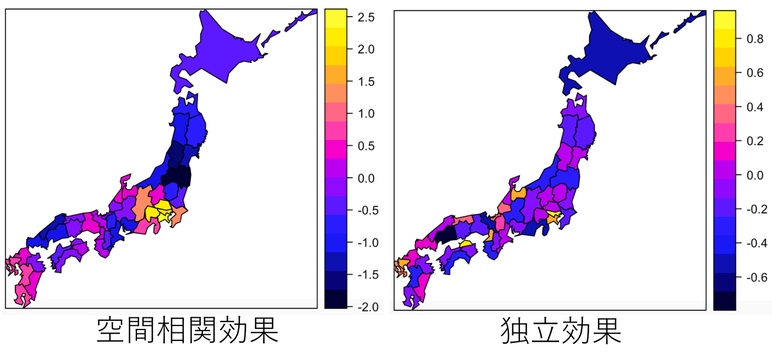
結果：説明変数からの効果

- 曜日の影響を除くと感染者数は滑らかに時間変化（左下図）
- 20～50代の労働世代の感染が多い（右図）。東京・大阪周辺では20代の陽性者数が顕著に多く、その他地方では40-50代が顕著に多いという推定も得られた(図は省略)
- 滞留人口が多く、かつ商業地に集中するほど陽性者数が増えるという結果も得た
- 所得は10%水準で正に有意。



結果：空間効果

- 都道府県別の空間相関成分と独立成分をそれぞれ推定した
- 空間相関成分: 説明変数の影響を差し引いても首都圏は高リスク
- 独立成分: 東京, 大阪, 富山, 香川などは局所的なホットスポット



大学共同利用機関法人 情報・システム研究機構
統計数理研究所
The Institute of Statistical Mathematics